# Outlier Detection in Large Radiological Datasets using UMAP

Mohammad Tariqul Islam Jason W. Fleischer

Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544

## TL;DR

- We introduce an unsupervised framework for detecting outliers in radiological image datasets.
- We analyze three public datasets: ChestX-ray14, CheXpert, and MURA.
- Our algorithm can discover erroneous and mislabeled x-rays and identify subgroups within them.

## Motivation

- Annotating and curating large radiological datasets is a difficult task.
- The annotations can have errors due to faulty perceptions, interpretations, and human errors. Existing tools may fail to check for signal quality.

We tackle these issues through clustering and visualizing the topology of the radiological dataset using dimensionality reduction, specifically the uniform manifold approximation and projection (UMAP) algorithm. **Outliers are different from the main data but can have similarities among themselves, which UMAP identifies.**

## Lateral X-rays in ChestX-ray14

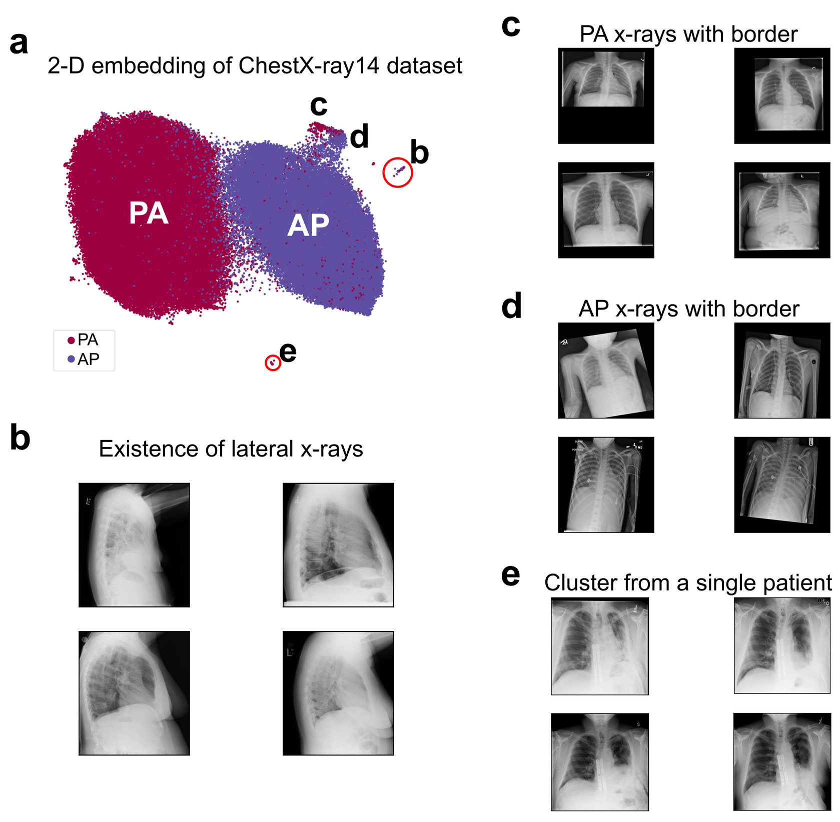


Figure 1. Outlier detection in the ChestX-ray14 dataset. (a) 2-D embedding. Labeled clusters from (a) are: (b) Lateral x-rays which were not supposed to be in the dataset, (c) PA x-rays with borders, (d) AP x-rays with borders, and (e) cluster from a single patient.

## Corrupted Images in CheXpert

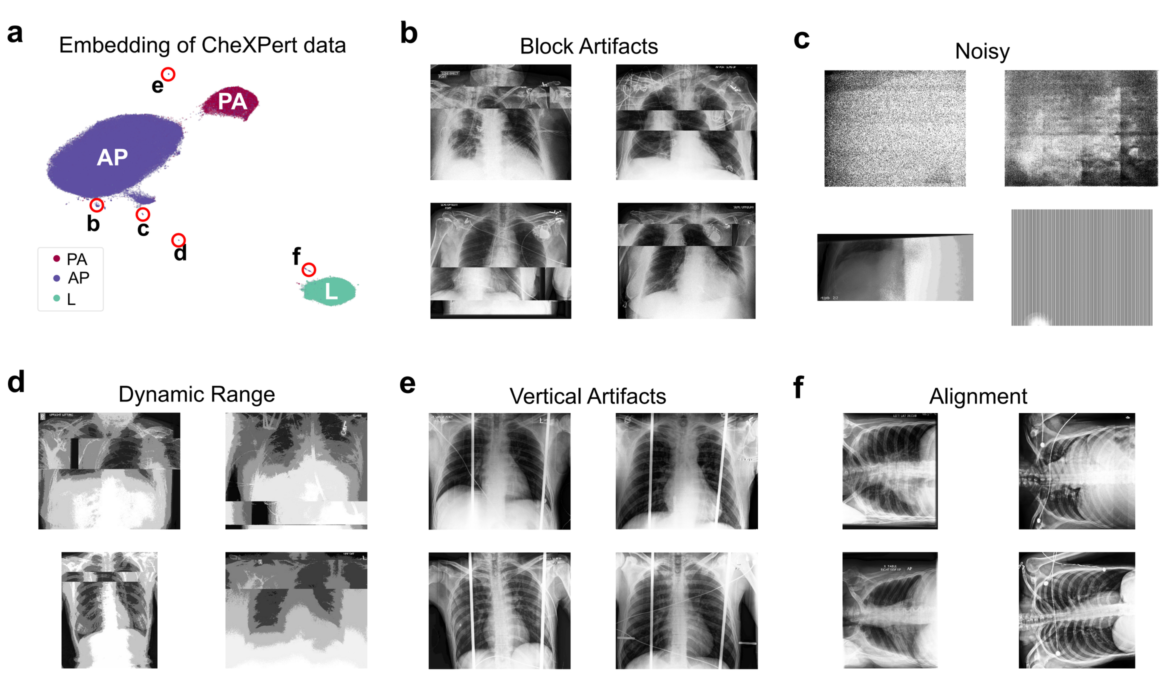


Figure 2. Outlier detection in the CheXpert dataset. (a) 2-D Embedding. Example images with (b) block artifacts, (c) noise, (d) improper dynamic range, (e) vertical artifacts, and (f) alignment issues.

## Comparison

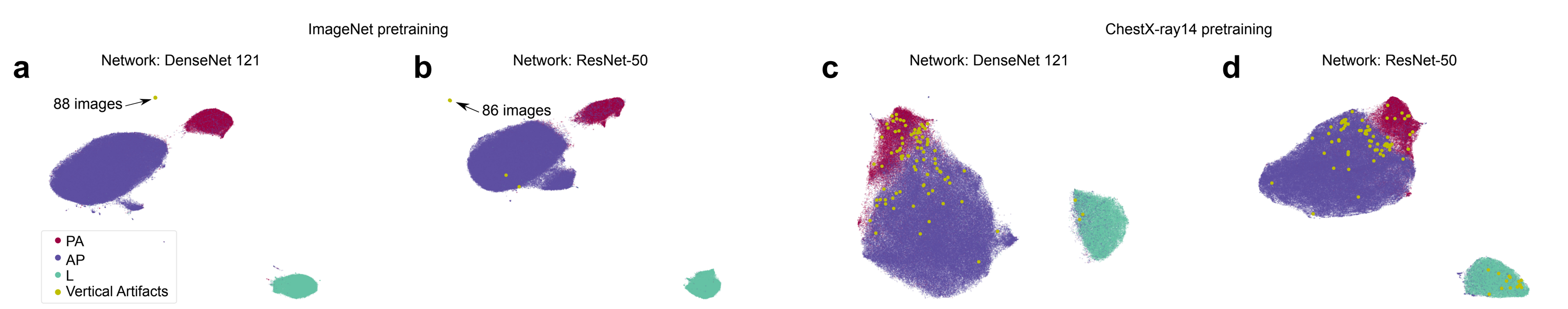


Figure 3. Embedding of CheXpert dataset using different pre-trained models. DenseNet-121 and ResNet-50 trained on ImageNet (left two) and ChestX-ray14 (right tow) datasets. Each yellow point represents an image with vertical artifact (from cluster e in Fig. 2 (a)) indicating chest x-ray pre-trained models fail to identify these as outliers.

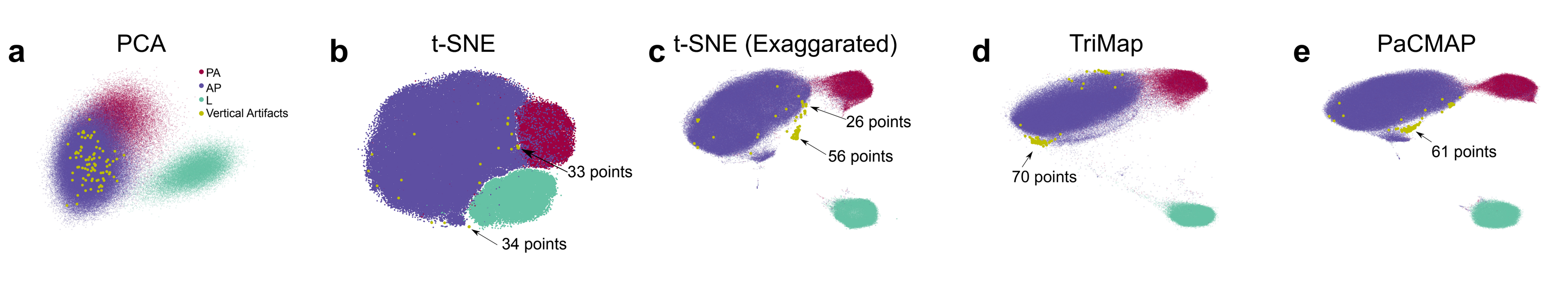


Figure 4. Embedding of CheXpert dataset using several dimensionality reduction algorithms. (a) PCA, (b) t-SNE, (c) t-SNE (exaggerated), (d) TriMap, and (e) PaCMAP. Each yellow point represents an image with vertical artifacts (from cluster e in Fig. 2 (a)).

## Outlier Detection Process

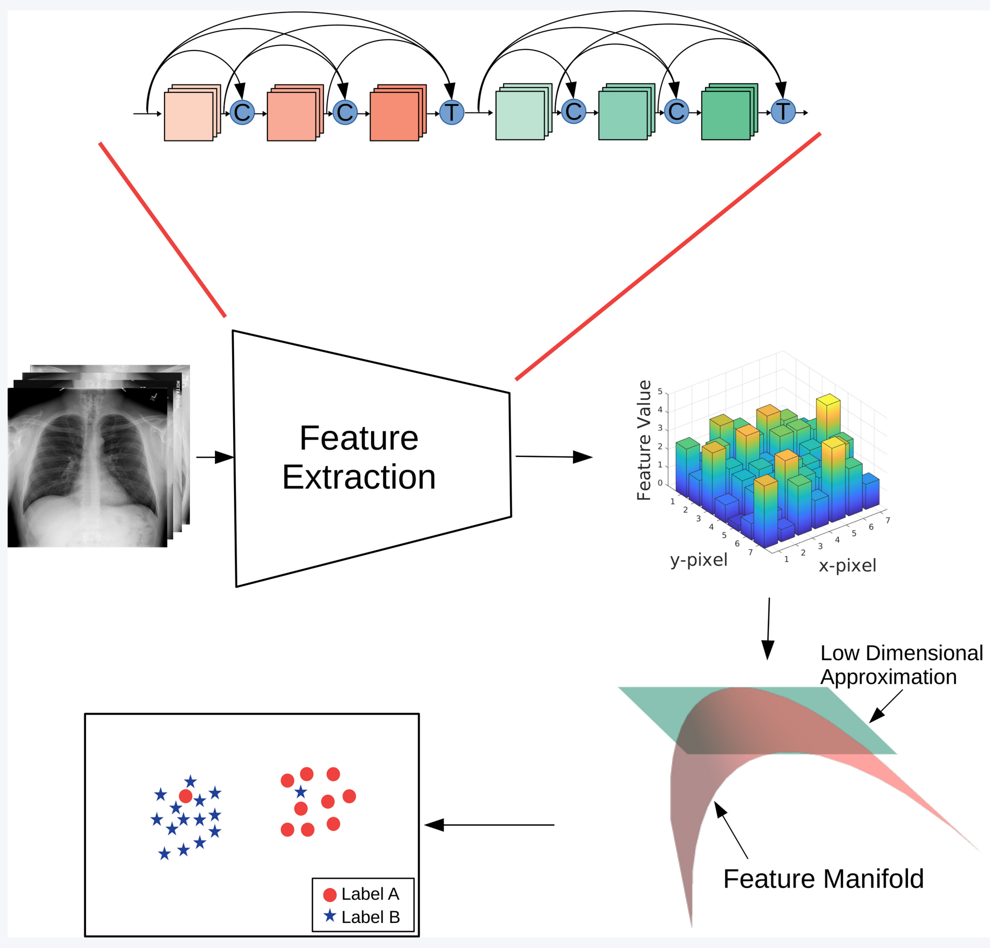


Figure 5. Schematic of the outlier search algorithm. Image features extracted from a DenseNet-121 neural network are projected onto a low-dimensional space (2-D plane) using UMAP.

- Major parts: feature extraction and dimensionality reduction.
- Feature extraction, using DenseNet-121 trained on ImageNet, makes the input to dimensionality reduction robust to image variabilities - varying resolution, different contrast, brightness, alignment, and registration issues.
- UMAP dimensionality-reduction method produces 2-D approximation of the high-dimensional features and clusters images with similar features together.

## Mislabeled X-rays in MURA

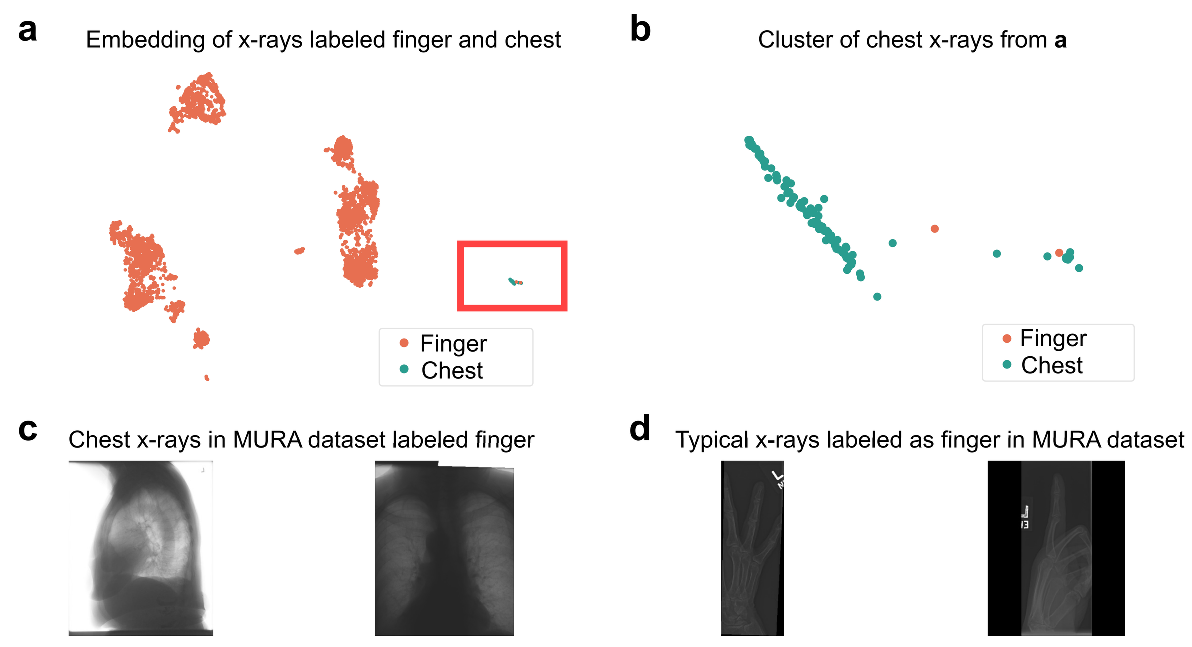


Figure 6. Embedding of 'finger' x-rays from MURA dataset and 100 chest x-rays from CheXpert dataset using UMAP. (a) Scatter plot of the embedding. The cluster of chest x-rays is marked using a red rectangle. (b) Scatter plot in the red rectangle. (c) two x-rays labeled 'finger' are actually chest x-rays. (d) Typical finger x-rays from the MURA dataset.

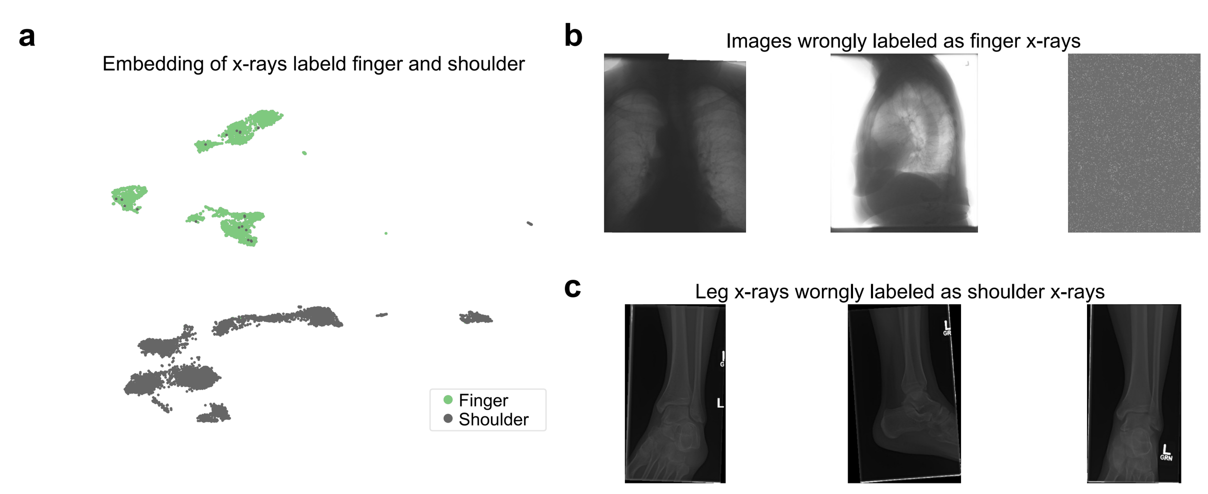


Figure 7. Embedding of 'finger' and 'shoulder' x-rays from MURA dataset using UMAP. (a) 2-D scatterplot of the embedding. (b) Chest x-ray and non-x-ray images were discovered which are labeled as 'finger' x-rays. (c) Leg x-rays labeled as 'shoulder' x-rays.

## Conclusions and Future

- UMAP can be an effective tool for summarizing datasets and identifying outlier images.
- We performed a retrospective analysis of large x-ray datasets, however, it will make strides during the initial assembly of a dataset.
- The methods are graph-based and agnostic to the underlying data type, and thus, can be extended for mixed modality.

## References

[1] Michael A Bruno, Eric A Walker, and Hani H Abujudeh. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*, 35(6):1668–1676, 2015.

[2] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[3] Stephen Waite, Jinel Scott, Brian Gale, Travis Fuchs, Srinivas Kolla, and Deborah Reede. Interpretive error in radiology. *American Journal of Roentgenology*, 208(4):739–749, 2017.

Acknowledgement: DataX Grant



jasonf@princeton.edu